Int. J. Appl. Math. Comput. Sci., 2023, Vol. 33, No. 1, 7–20
DOI: 10.34768/amcs-2023-0001

EFFICIENT FACE DETECTION BASED CROWD DENSITY ESTIMATION USING CONVOLUTIONAL NEURAL NETWORKS AND AN IMPROVED SLIDING WINDOW STRATEGY

ROUHOLLAH KIAN ARA [a,*], ANDRZEJ MATIOLANSKI [a], MICHAŁ GREGA [a], ANDRZEJ DZIECH [a], REMIGIUSZ BARAN [b]

[a]Institute of Telecommunications
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
e-mail: rouhollah.kian.ara@ict.agh.edu.pl,
{andrzej.matiolanski,michal.grega,andrzej.dziech}@agh.edu.pl

[b]Department of Computer Science, Electronics and Electrical Engineering
Kielce University of Technology
ul. Żeromskiego 5, 25-369 Kielce, Poland
e-mail: r.baran@tu.kielce.pl

Counting and detecting occluded faces in a crowd is a challenging task in computer vision. In this paper, we propose a new approach to face detection-based crowd estimation under significant occlusion and head posture variations. Most state-of-the-art face detectors cannot detect excessively occluded faces. To address the problem, an improved approach to training various detectors is described. To obtain a reasonable evaluation of our solution, we trained and tested the model on our substantially occluded data set. The dataset contains images with up to 90 degrees out-of-plane rotation and faces with 25%, 50%, and 75% occlusion levels. In this study, we trained the proposed model on 48,000 images obtained from our dataset consisting of 19 crowd scenes. To evaluate the model, we used 109 images with face counts ranging from 21 to 905 and with an average of 145 individuals per image. Detecting faces in crowded scenes with the underlying challenges cannot be addressed using a single face detection method. Therefore, a robust method for counting visible faces in a crowd is proposed by combining different traditional machine learning and convolutional neural network algorithms. Utilizing a network based on the VGGNet architecture, the proposed algorithm outperforms various state-of-the-art algorithms in detecting faces ‘in-the-wild’. In addition, the performance of the proposed approach is evaluated on publicly available datasets containing in-plane/out-of-plane rotation images as well as images with various lighting changes. The proposed approach achieved similar or higher accuracy.

Keywords: crowd density, face detection, head pose variations, various lighting conditions, occlusion.

1. Introduction

1.1. Crowd density estimation. Social advances and the great increase in the world population had led to an increasing number of social events, such as concerts, sports events, pilgrimage, political rallies, etc. All of those assemblies may lead to public security problems resulting in mass panic, riots, or violent protests. Therefore, the use of automated or semi-automatic crowd control techniques such as queue management, detecting and counting people, and estimating crowd flows play an essential role in human safety. Overall, crowd density estimation and people counting problems are categorized into direct and indirect approaches (Conte *et al.*, 2010).

The direct approaches (object/face detection based) attempt to segment and detect each person in crowd scenes and count them using effective detectors/classifiers (Zhao *et al.*, 2008). Detecting people using this method becomes more difficult in a high-density crowd where occlusions occur (Kotan *et al.*, 2021). Nevertheless, crowd counting under substantial occlusion problems has

[*]Corresponding author

been addressed adapting part-based detectors, such as a head detector (Sheng-Fuu *et al.*, 2001) or a pedestrian detector (Khatoon *et al.*, 2012). Still, the issue of counting people in the high-density crowd and occluded scenes remain unresolved.

In indirect approaches (feature-based), the various local features of the entire crowd are extracted (Saleh *et al.*, 2015). However, this approach is flawed in various crowded scenes or under perspective distortions. Also, it does not propose a solution for tracking faces. Solutions for the aforementioned shortcomings are listed below.

Ma *et al.* (2004) utilized geometric correction (GC) to convert all objects in the scene from various scales to a common scale to address perspective issues in the indirect approach. Fradi and Dugelay (2012) proposed applying a perspective map normalization to weight all extracted features to compensate for variations in distance and density. In indirect approaches, the occlusions problem has been addressed by employing additional features, such as an edge histogram (Kong *et al.*, 2005), edge counting (Davies *et al.*, 1995), etc.

In this paper, we propose a model that addresses the unresolved occlusion issues in the direct approaches, as well as perspective issues caused by the indirect approaches. Moreover, the occlusion and pose variation issues have been addressed in the training phase of our neural network model. In addition, perspective and detection issues have been solved using the sliding window approach, which is specified and developed for use only in face detection in crowded approaches.

1.2. Face detection.

Detecting human faces in a crowd scene is a fundamental problem in computer vision. Significant progress has been made after seminal work by Viola and Jones (2004). The primary task of modern face detectors for real-world applications is to identify whether there is a frontal face in an image. Recent research in this area focuses on face detection challenges, such as variations in scale, head pose, exaggerated facial expression, significant occlusion, and various lighting conditions.

In recent years, deep learning architectures, such as convolutional neural networks (CNNs), have become the most popular solution for face detection as a special type of the object detection task in computer vision. With respect to performance it has overshadowed classical computer vision approaches.

In contrast to the traditional machine learning (ML) algorithms, a CNN extracts crucial and meaningful features, such as edges, in their first layers and combines them in order to detect shapes in the next layers. Further on, in fully connected layers, the CNNs learn how to utilize all these features in order to detect or classify objects in the scene.

Although CNNs are computationally expensive and generally require more data for training, compared with ML algorithms, they are effective at detecting the most complex features of images (Zitouni and Śluzek, 2022).

1.3. Counting people's faces in crowds.

Detecting faces in crowded scenes with the underlying challenges cannot be addressed using a single face detection method. In this paper a robust method of face counting in the crowd is proposed, that combines various traditional ML and state-of-the-art CNN algorithms.

First, we employ well-known traditional machine learning algorithms such as the histogram of oriented gradients (HOG) with support vector machines (SVMs) and the Haar feature-based cascade classifier to obtain the size of faces in the crowd images. Afterward, the obtained sizes are utilized to specify the step size between the window patches. Finally, the CNN is employed to classify and count faces in each patch. The sum of faces in the small patches gives the number of all detected faces in the entire crowd image. The details of this procedure are provided later in this paper.

We evaluate the proposed approach on our crowd image dataset (AGH Crowd Density Estimation Database—ACD) (Kian Ara and Matiolanski, 2019), which contains images with heavy occlusion and head pose variations.

Afterward, it is evaluated against the Pointing'04 dataset (Gourier *et al.*, 2004), which contains in-plane/out-of-plane rotations, and against the FEI dataset (Thomaz and Giraldi, 2010), which contains images with very high lighting changes.

The major contributions of this study are as follows:

- designing a face detection based crowd density estimation approach using a developed sliding window strategy and a CNN model;
- proving our custom training dataset based on model requirements;
- proposing a data annotation strategy in order to overcome false detections caused by occluded and out-of-plane rotation face images;
- strengthening the model detection capability in a very dark environment by applying a new strategy.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 describes the proposed approach to face detection. Section 4 discusses our experiments and results. Lastly, Section 5 concludes this paper.

2. Related work

Face detection and recognition have evolved as one of the most widely used biometric techniques in many areas, such as public security, finance, crowd management, and safety in recent years.

The work by Viola and Jones (2004) was the first face detection framework that proposed rectangular Haar features in a cascaded Adaboost classifier to detect the faces in real-time with promising accuracy and efficiency. Its drawbacks such as large feature sizes and the incapability to detect faces in the wild motivated the researchers to address the problems with more complicated features like HOG (Zhu *et al.*, 2006), scale-invariant feature transform (SIFT), and speeded up robust features (SURF) (Li and Zhang, 2013).

Another approach to improve the accuracy of detectors was to separately train several models for various head poses and scenes (Jones and Viola, 2003; Li *et al.*, 2002). Chen *et al.* (2014) introduced a model to perform face detection combined with face alignment in order to improve the accuracy and speed of the object detector in challenging datasets.

In recent years, deep CNNs have achieved remarkable success in face detection applications. Jiang and Learned-Miller (2017) investigated Faster R-CNN (Ren *et al.*, 2015), a state-of-the-art object detector, and achieved a significant performance increase in both speed and accuracy. Combining Faster R-CNN with hard negative mining and ResNet, Wan *et al.* (2016) achieved remarkable performance on the FDDB face detection benchmark. Sun *et al.* (2017) improved the performance of the Faster R-CNN algorithm by proposing several strategies such as feature contention, multi-scale training, and hard negative mining. Although many studies focused on occluded face detection (cf., e.g., Mahbub *et al.*, 2016; Opitz *et al.*, 2016; Yang *et al.*, 2015) the performance of face detectors is imperfect when the faces are severely occluded and the poses vary. In the next section, we will describe our approach in more detail. This approach has the potential to overcome the limitations of the existing studies.

3. Proposed method

Many algorithms have been evaluated on low-density crowds, such as UCSD, Mall, PETS datasets with a density of 11–46, 13–53, 3–40 people per image respectively (Chan *et al.*, 2008; Chen *et al.*, 2012; Ferryman and Ellis, 2010). Other researchers have principally focused on counting people in extremely dense crowd images. In such images, each individual may occupy a few pixels only (Idrees *et al.*, 2013). The proposed face-based algorithm in this study has been evaluated on medium-density crowds, with a density of 21 to 905 faces per image. To obtain a reasonable evaluation of our solution, we trained and tested our model on our substantially occluded dataset (AGH Crowd Density Estimation Database, or ACD) (Kian Ara and Matiolanski, 2019).

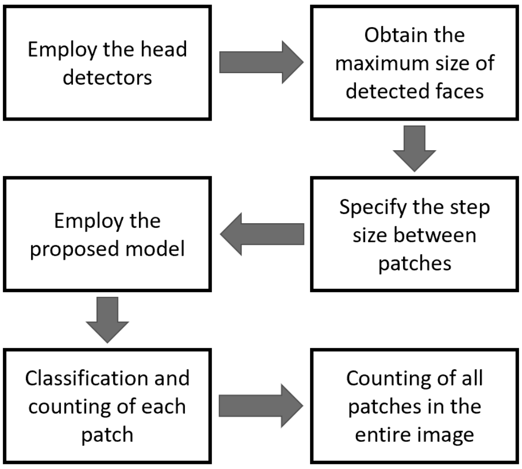

Employ the head detectors
Obtain the maximum size of detected faces
Specify the step size between patches
Employ the proposed model
Classification and counting of each patch
Counting of all patches in the entire image

Fig. 1. Flowchart of the proposed algorithm.

Our goal is to estimate the number of people based on faces in crowd scenes under significant occlusion. To address the problem of occlusion and variation in poses, we introduce occlusion during the training of the model. By segmenting images into several random sections in the training procedure of the CNN algorithm, the model learns various features of faces and can mitigate the negative effects of occlusion and variations in the pose.

In contrast to many state-of-the-art detection and classification approaches, which utilize windows of varied sizes and time-consuming pyramid structures in object detection (Liu *et al.*, 2012; You-jia and Jian-wei, 2010; Han *et al.*, 2014),we propose an improved sliding window technique for use in crowd image analysis. First, we divide the images into 50 by 50 patches (pixels) and apply the classifier (CNN) to each of them. The sum of the faces detected in each patch makes the final estimate. To avoid multiple counts of single faces that may appear at the border between patches, we used a combination of several approaches, such as Haar-cascade detection and the SVM classifier with HOG, to obtain the size of each head.

A new sliding window technique is proposed to determine the step size between patches based on the information (size of the whole face) obtained by the face detectors in the first stage. The flowchart of the proposed algorithm is shown in Fig. 1.

3.1. Data preparation.

We trained and tested the proposed model on our dataset (Kian Ara and Matiolanski, 2019). It consists of 48,000 patches of head images obtained from the crowd scenes for training purposes and 109 crowd scenes for testing the model performance. The total number of unique faces in the training dataset images is 2181 (48,000 cropped from various aspects of the face) and 16,501 in the test dataset images.

Images were manually labeled before feeding the dataset to the model. First, the database images are converted to grayscale, and the position of the faces in each image is determined by marking them with a single red dot on each face. To train on the most significant and meaningful part of faces, red dots are placed in the center of the nose and eyes triangle.

Subtracting images with red dots from grayscale images automatically yields the count of all labels in each image. Furthermore, these images were randomly segmented to generate 40,000 training samples of four classes:

- Class 0: negative (patches without any faces)
- Class 1: positive (patches consisting of one face)
- Class 2: positive (patches consisting of two faces)
- Class 3: positive (patches consisting of three faces)

As shown in Fig. 2, a new dataset is created by randomly selecting rectangles around this reference point (face center point). Using this strategy, hundreds of new data are created from just one face image. For instance, the main rectangle creates a full-sized face containing, eyes, a nose, and a mouth. Another rectangle could be an image containing only the nose and one eye. If the lower or upper edge of the rectangle is at the center point, the most covered face image is created, with almost 25% of the entire face visible. Therefore, using this data preparation method, we can create a model that can recognize images of masked faces up to 3/4 of the total covered. Further on, in Section 3.3, we shall show how discarding the face area with less than 25% visibility (which we claim not to detect) helps us not to double-count the faces in the margin between neighboring windows in the sliding window strategy.

Training the model with specified segmented regions (the region of interest includes at least one of the eyes or nose) enables the CNN model to observe and learn facial features from different aspects and focuses on the most crucial parts of the face. This alleviates the effect of occlusion and ensures that the model is robust against head posture variation.

3.2. Convolutional neural network (CNN). Our training dataset consists of 40,000 face images in 4 classes. We train the CNN to recognize and classify each of these classes. The CNN used in this study is based on the main architecture of the state-of-the-art VGG network (Simonyan and Zisserman, 2014) and is inspired by Rosebrock (2021). The VGGNet has been trained on ImageNet, the most comprehensive hand-annotated visual dataset (Simonyan and Zisserman, 2014). Among the top-performing CNN models, VGG is notable for its simplicity and uniform architecture. In contrast to the 16-layer VGG network, with a 7.3% error rate, there are much more complex models such as Microsoft's ResNet model with a 3.6% error rate, but many more layers (152 layers) (Simonyan and Zisserman, 2014) compared with the VGG.

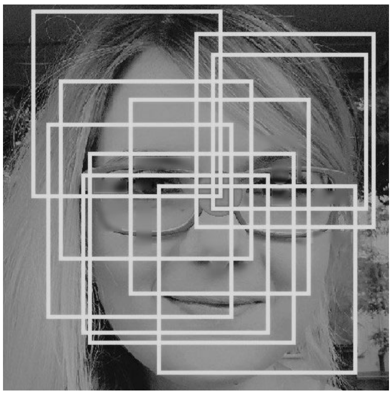


Fig. 2. Generating dataset/images for training the CNN model.

Notwithstanding its benefits, the VGG neural network has a few drawbacks. It is very slow in the training phase and its weights are extremely large, making it a time-consuming task to deploy the VGG.

The proposed CNN model contains 5 convolutional layers, 3 max-pooling layers, and a set of fully connected layers. The first convolutional layer consists of 32 kernels of size 3×3 applied with a stride of 1 and padding is set to "same".

The CNN models require nonlinear properties in order to represent more sophisticated features that are more adaptable to the real world.

A new feature map was created in the first layer by applying a nonlinear activation function, namely ReLU (rectified linear unit). The monotonicity, continuity, and scale invariance are the most important properties of the Relu activation function needed in the optimization process. This function is defined as

$$\mathrm{ReLU}(x) = \max(0, x), \tag{1}$$

where the output of ReLu is mapped between zero (the absence of the feature) and the input value x.

In order to reduce the internal covariate shift of the network, the batch normalization technique was applied between the layers of the CNN model (Ioffe and Szegedy, 2015).

The max-pooling is performed on a 3×3 pixel window to reduce the resolution of the feature map while preserving active features. The most popular regularization technique, dropout (with the probability set to 25%), is applied to the hidden layer nodes in order to reduce the effect of overfitting caused by model fitting on the training data. The number of filters has been increased from 32 to 64 in the second block of the model structure to learn the most complex facial features. At this stage, 25% dropout is applied again.

The next block of the proposed model is similar to the second block, except that the number of filters is increased to 128 cores.

Finally, to attain predicted probabilities for each class label (label 0 = negative image, label 1 = one face, label 2 = two faces, label 3 = three faces), fully-connected layers with a dropout of 50% of nodes are followed by a Softmax transformation (classifier) at the end of the model structure. We aim to classify into four categories, hence the categorical cross-entropy loss function (Softmax Loss = Softmax activation + Cross-Entropy loss) is used to distinguish between estimated and ground truth labels. The CNN aims to optimize its weights and biases by minimizing the cross-entropy error through the cost function.

To avoid overfitting, another regularization technique, data augmentation, is employed in the training phase. Data augmentation is a technique that allows generating more training data from the existing dataset by applying random transformations, such as rotations, variations in the brightness of images, horizontal/vertical shifts, horizontal/vertical flips, etc. (Fig. 3). The rest of the model's hyperparameters are listed as follows: The total number of training epochs = 300, the initial learning rate for the Adam optimizer = 0.001, and the batch size = 32. The proposed CNN is implemented using Keras, an open-source deep-learning library (Keras, 2015).

The proposed model is trained on 80% of the randomly selected images and afterward tested on the 20% of subsequent images. As illustrated in Fig. 4, our CNN obtained 84.41% classification accuracy on the training set and 84.36% accuracy on the testing set. The architecture of the proposed CNN is presented in Table 1. The CNN uses tensors of input shape (image height, image width, color channels), ((50, 50, 3) in this study). Additionally, a 3D tensor shape (height, width, channels) is also the result of any Conv2D and MaxPooling2D layer operations. As shown in the table, applying more layers to the CNN architecture results in a smaller output shape of the subsequent layer.

In order to increase the classification accuracy of the model, 8000 images with a classification confidence score greater than 0.95 are appended to the initial dataset. Afterward, the proposed model was retrained on the new dataset which improved the classification accuracy by 3%. The model accuracy for the binary classifier was improved up to 97.58% and was subsequently used to compare the proposed model with state-of-the-art face detectors.

Fig. 3. Data augmentation example.

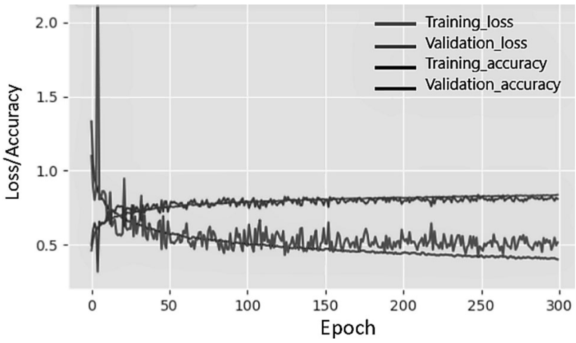


Fig. 4. Training loss and accuracy of the model.

3.3. Sliding-window and occlusion solutions. In order to address the problem of occlusion and variation in head poses, we include occlusion in the training phase of the model. The proposed model is trained on the images of faces with an occluded area up to 75% where faces are 50% occluded in both vertical and horizontal directions. Figure 5 demonstrates how the proposed method avoids counting faces multiple times (twice or more), where a part of the face can appear in more than one patch.

To estimate the number of people in crowded scenes, images are divided into patches of 50×50 pixels. These patches are then individually classified/labeled by the CNN model. Counting the classification result of all patches gives the final prediction result.

We propose a modified sliding window technique to avoid errors caused by the appearance of face parts in two or more patches.

In the proposed technique fixed-size windows are shifted from left to right and from top to bottom to identify faces using the CNN classifier (see Fig. 6). In this method, first, the size of the detected head is specified by various approaches, and afterward, the step size between each window is calculated according to the obtained dimensions as

$$\Delta = L + \frac{M}{4}, \tag{2}$$

where L = 50 (pixels in vertical and horizontal directions)

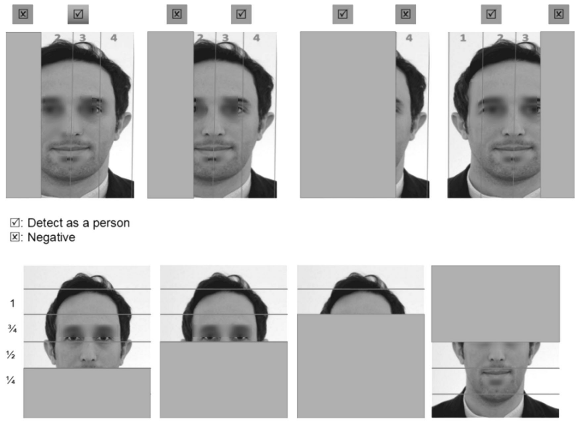


Fig. 5. Proposed algorithm recognizes faces with an occluded area of up to 75%.

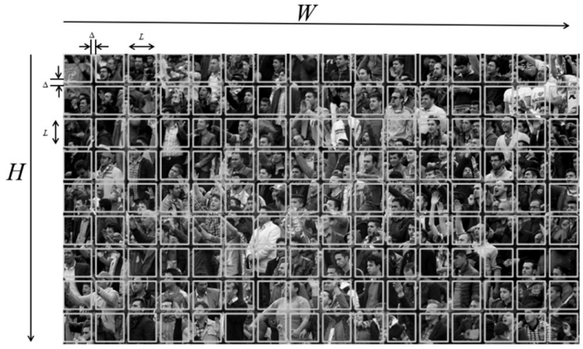


Fig. 6. Scene-scanning (start from the top-left corner, shift by delta pixels).

is the length of the window and M is the maximum length of the detected face (pixels) with various approaches such as Haar-cascade detection, SVM classifier with HOG features, etc.

As mentioned before, our model is trained to recognize faces that are occluded up to 75% of their area (Fig. 5). While this is useful in detecting occluded faces it can also cause our algorithm to count faces that are split between neighboring windows twice. In order to overcome this issue, we have included a margin between neighboring windows. This margin spans horizontally (for 25% of window width) and vertically (for 25% of window height).

Table 1. Summary of the CNN architecture used in our experiments.

Layer (type)	Output shape	Param #
Conv2D	(None,48,48,32)	896
Activation	(None,48,48,32)	0
Batch_Normalization	(None,48,48,32)	128
Max_Pooling2D	(None,16,16,32)	0
Dropout	(None,16,16,32)	0
Conv2D	(None,16,16,64)	18496
Activation	(None,16,16,64)	0
Batch_Normalization	(None,16,16,64)	256
Conv2D	(None,16,16,64)	36928
Activation	(None,16,16,64)	0
Batch_Normalization	(None,16,16,64)	256
Max_Pooling2D	(None,8,8,64)	0
Dropout	(None,8,8,64)	0
Conv2D	(None,8,8,128)	73856
Activation	(None,8,8,128)	0
Batch_Normalization	(None,8,8,128)	512
Conv2D	(None,8,8,128)	147584
Activation	(None,8,8,128)	0
Batch_Normalization	(None,8,8,128)	512
Max_Pooling2D	(None,4,4,128)	0
Dropout	(None,4,4,128)	0
Flatten	(None,2048)	0
Dense	(None,1024)	2098176
Activation	(None,1024)	0
Batch_Normalization	(None,1024)	4096
Dropout	(None,1024)	0
Dense	(None,4)	4100
Activation	(None,4)	0
Total Pramas		2,385,796
Trainable Pramas		2,382,916
Non-Trainable Pramas		2,880

3.4. Histogram equalizations as a lighting solution.

Histogram equalization (HE) is a broadly used technique for image contrast enhancement due to its simplicity and reasonable performance, particularly in a condition, where the image is too dark or too bright and contrast enhancement of the entire input image is required (Fig. 7). HE aims to convert the given distribution to a uniform distribution assuming the pixel values lie between 0 and 255 and it is calculated as follows:

$$S_k = T(r_k) = \sum_{j=0}^{k} p(r_j) = \frac{L-1}{MN} \sum_{j=0}^{k} n_j, \qquad (3)$$

where $k \in [0, \dots, L-1]$ for an n bit image, $L = 2^n$, S_K, r_k, P, r_j, L, MN, n stand for the number of pixels in the equalized image, the number of new frequencies, the total frequency that corresponds to a specific value of r_j, the range of values from 0 to $L-1$, the maximum intensity value, the total number of pixels in the image, and the number of pixels with intensity j, respectively.

First, we added 200 single-face images to our ACD (Kian Ara and Matiolanski, 2019) and applied our algorithm in the two-class estimator mode on the Pointing'04 dataset (Gourier *et al.*, 2004) and compared the performance results with various algorithms. The FEI dataset (Thomaz and Giraldi, 2010) contains 2800

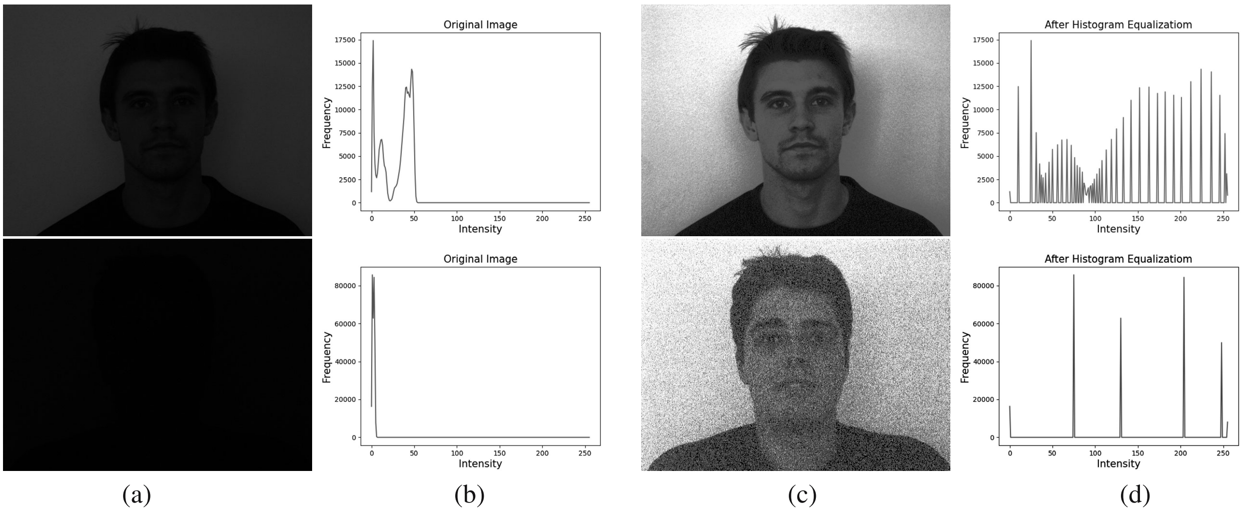


Fig. 7. Original image (a), intensity histogram of the original image (b), image after histogram equalization (c), intensity histogram of the equalized image (d).

face images including dark (faces are recognizable to the human eye), very dark (faces are not recognized by a human eye), excessively dark (faces are invisible in the images), and out-of-plane rotation images (Table 2).

We investigated applying HE to all images and found that despite its ability to reconstruct hidden faces in an image, it had a few drawbacks. In high-quality images, the image quality decreases due to the HE average shift approach. Moreover, the computation power requirements increase significantly. In order to reconstruct faces from the dark images, we proposed an algorithm depicted in Fig. 8.

Initially, we need to identify extremely dark images where faces are invisible and divide our dataset into two categories based on their brightness intensities. For noise removal, we first applied a Gaussian filter, and afterward calculated the average luminance intensity of the entire image. Empirically, we have found that images with an average brightness intensity of less than 15 indicate intense dark images. A 5 $\times$ 5 Gaussian kernel with a default border type is used to smooth the image noise. The formula for a Gaussian function in two dimensions is (Shapiro and Stockman, 2001)

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\Big(-\frac{x^2 + y^2}{2\sigma^2}\Big), \qquad (4)$$

where σ is the standard deviation of the Gaussian distribution, x and y are the distances of the pixels from the origin in the horizontal and vertical axes, respectively.

After applying HE, the non-local means denoising technique (Buades *et al.*, 2011) is utilized to reduce the noise effect generated by the HE algorithm in reconstructed images. The suggested algorithm chooses a pixel and dedicates a window around it (7 $\times$ 7 in this study), then it looks for similar patches in order to replace that specified pixel value with the average of all similar windows. It implements this process over the entire image pixels, which makes the process of denoising images time-consuming compared with the other denoising methods, but with better performance.

4. Experiments and evaluation

As mentioned before, we collected 109 images with a person count ranging between 21 and 905 with an average of 151 individuals per image from various sources. Some of the examples of images with the associated ground truth count are illustrated in Fig. 9. Figure 10 demonstrates a few examples of our model performance in each subwindow.

4.1. Comparison of methods for pose variations, severe occlusion, and various lighting conditions. Studies involving face detection differ in objects and training and testing datasets. Therefore, an accurate comparison of this work with these artifacts may not be entirely meaningful.

Therefore, in this study, we separately compare solutions for face detection with pose variations (such as rotation of faces from -90 to 90 degrees), detection under various occlusion levels (such as faces with 25%, 50%, and 75% occlusion), and in images with various lighting conditions.

In order to evaluate the comparison of different methods in real-world applications, we utilize our dataset (AGH Crowd Density Estimation Database, or ACD), which is a more challenging benchmark compared with the existing datasets. In this study, the sum of absolute differences (SAD) and the mean of absolute errors (MAE) were utilized to evaluate the performance of various face

Table 2. Summary of the datasets.

Conditions/Dataset	ACD	Pointing'04	FEI dataset
Number of images	109	1302	2800
Number of faces	16501	1302	2800
Head pose variations	Out-of-plane	In/out-of-plane	Out-of-plane
Lighting condition	Normal	Normal	Normal/dark/very dark
Occlusion	Up to 75 %	Normal	Few faces covered by glasses

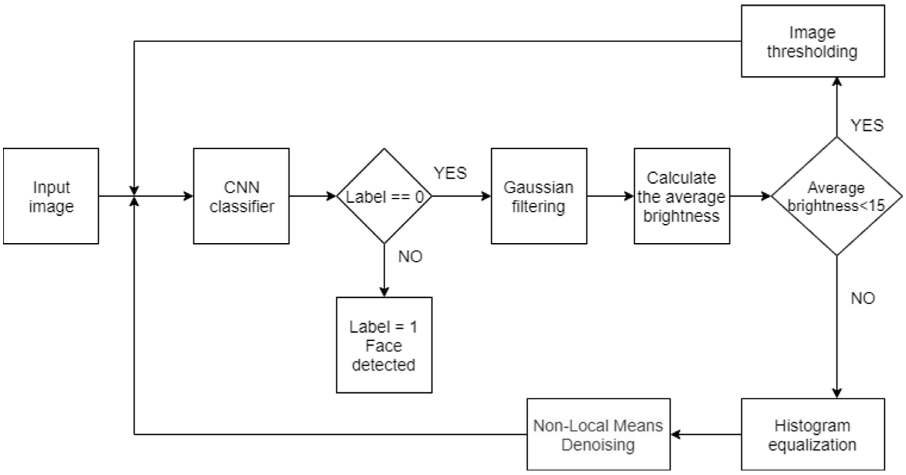


Fig. 8. Flowchart of implementation of the proposed face detection algorithm on dark images.

detectors in crowds using the following formulas:

$$\mathrm{SAD} = \sum_{i=1}^{N} |y_i - x_i|, \tag{5}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - x_i|, \tag{6}$$

where N is the number of test samples, y and x are the estimated face count and ground truth, respectively.

Recall or sensitivity (the ratio of correctly predicted positive observations to all true observations) was used to assess the face detection model performance. It can be the model metric when the cost of a false negative (FN) or Type II error is high,

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{7}$$

where TP and FN, denote the numbers of true positives and false negatives, respectively.

The proposed method is evaluated in two scenarios. In the first scenario, an empirical comparison of the performance of the proposed approach and the well-known face detection libraries and toolkits such as the Haar cascade face detector, the deep learning face detector in OpenCV (DNN) (Open CV, 2019), and the HOG with a linear SVM and a maximum-margin object detector (MMOD) with CNN based features in Dlib (King, 2009), is compared in the most challenging conditions. The capabilities of these algorithms on an example image are demonstrated in Table 3.

The algorithms are trained on images with specific sizes so an accurate comparison of these methods may not be possible. Therefore, the algorithms are evaluated according to their desired image sizes. As is shown in Table 3, the proposed approach can recognize faces in various poses and faces with up to 75% coverage. The DNN and MMOD methods outperform the Haar, LBP, and HOG-based face detectors, especially in pose variations. HOG-based detectors are more accurate than LBP and Haar cascades, with fewer false positives (FPs). However, FP is the major drawback of the Haar cascade face detector appropriate thresholding is used to overcome this problem. In this study, we overcome false-positive errors by discarding faces larger than twice the average perceived face size. HOG and Haar feature-based cascade classifiers are less robust to occlusion, which can be utilized in the first phase of the proposed approach, where we only want to obtain the full size of the human face in a crowd.

In the second scenario, the proposed method is evaluated on the Pointing'04 dataset (Gourier *et al.*, 2004) and the FEI dataset (Thomaz and Giraldi, 2010). The Pointing'04 dataset contains 1302 large-sized (faces with a minimum size of 140 × 140) single faces with various rotations. The FEI dataset contains 2800

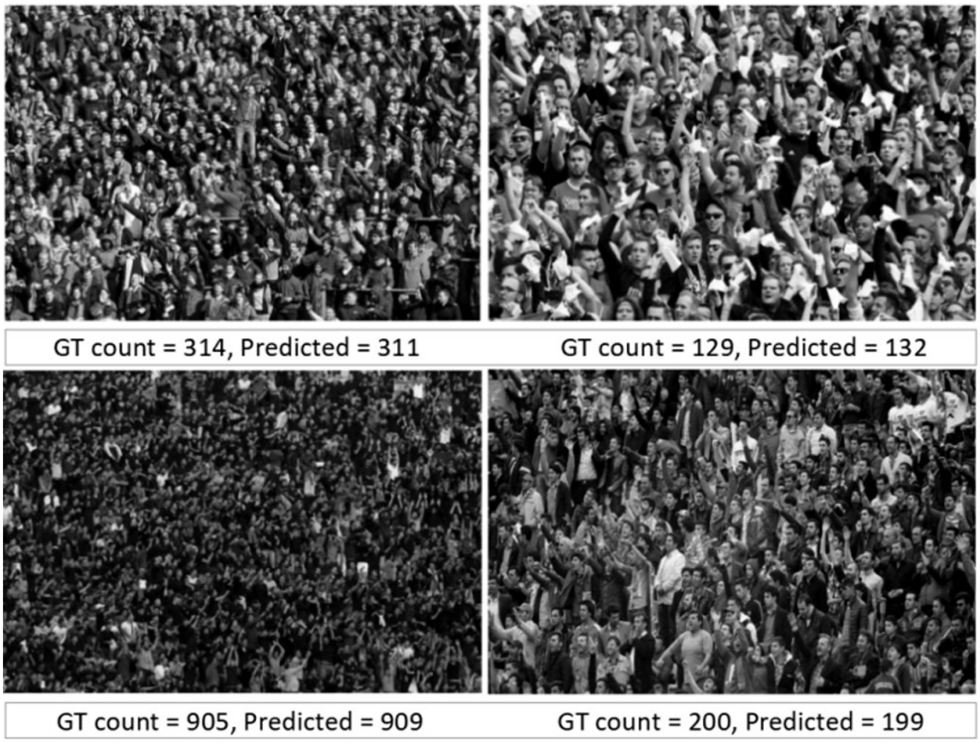

GT count = 314, Predicted = 311
GT count = 129, Predicted = 132
GT count = 905, Predicted = 909
GT count = 200, Predicted = 199

Fig. 9. Estimation results of the proposed method.

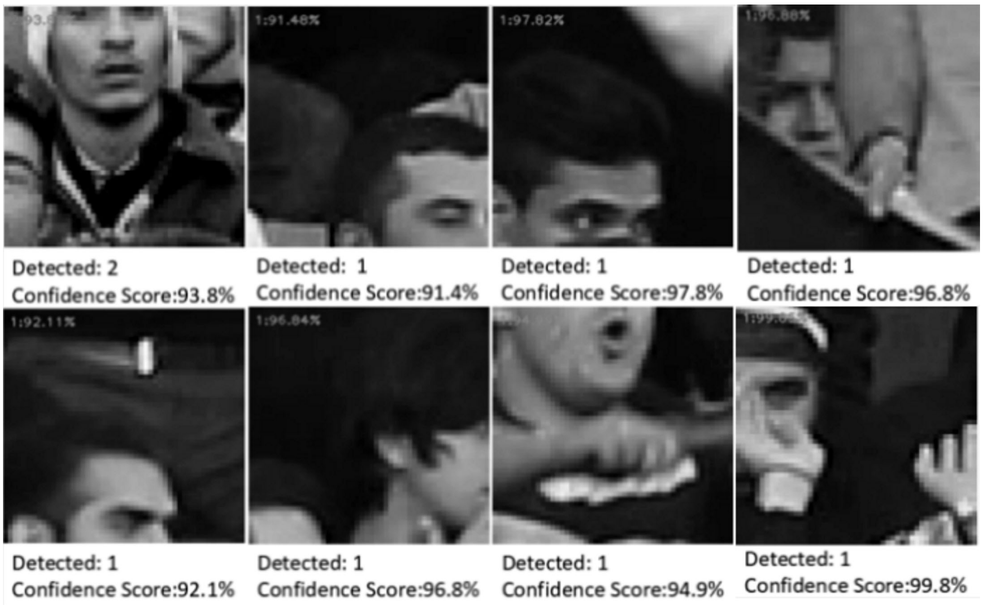
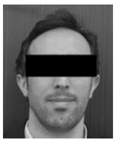
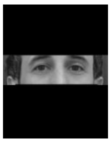
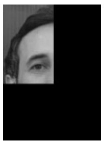
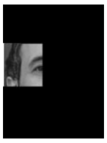
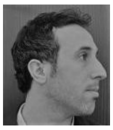
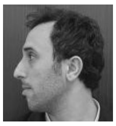
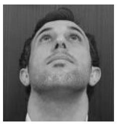
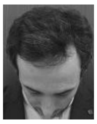


Fig. 10. Face detection result of the proposed method on each sub-window. The evaluation was examined on 50×50 images; the displayed images are resized to 400×400 pixel images.

Table 3. Successful detections with head pose variations are indicated by ‘Yes’.

								
Haar	No	No	No	No	No	No	Yes	No
LBP	Yes	No	No	No	No	No	Yes	No
DNN	Yes	No	No	No	Yes	Yes	Yes	No
HOG-SVM	Yes	No	No	No	No	No	Yes	No
MMOD	Yes	No	No	No	Yes	Yes	Yes	No
Proposed	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

single faces photographed in various lighting conditions. The performance of the proposed algorithm has been compared with state-of-the-art face detectors, some of which are presented by Tsai *et al.* (2018).

These algorithms are Adaboost (Viola and Jones, 2004), Simple (Pai *et al.*, 2006), the head detector (Marin-Jimenez *et al.*, 2011), the object detector (Cevikalp and Triggs, 2012), the tree-structured part model (Face p146) (Zhu and Ramanan, 2012), the fast face detector (FFD) (Yang *et al.*, 2014), the hierarchical part model (HPM) (Ghiasi and Fowlkes, 2014), the deformable part model (DPM) (Orozco *et al.*, 2015), image quality assessment (IQA) (Chen *et al.*, 2015), Face++ (Face++, 2015), the maximum-margin object detector (MMOD) with CNN based features (King, 2009; 2015), normalized pixel difference (NPD) (Liao *et al.*, 2016), the single-stage headless (SSH) face detector (Najibi *et al.*, 2017) and the in-plane/out-of-plane face detector (IOP) (Tsai *et al.*, 2018), as well as the deep neural networks (DNN) module (Open CV, 2019). Comparative results are exhibited in Table 4.

Notwithstanding that our model was prepared and trained for small-size faces (images consisting of faces between 18 and 50 pixels) in a crowded environment, it achieves prominent performance on both datasets.

4.2. Results by the proposed method. In this section, a combination of different methods is compared and presented in Table 5. For comparison, we used 109 images with 21 to 905 individuals per image. The size of each face in the images varies between 18×18 and 162×162 pixels.

The first row in Table 5 shows the results of using only the Haar cascade to specify the size of the faces, yielding 1237 SAD and 11.34 MAE. The second row shows that using HOG with SVM interestingly reduces MAE to 6.2. A combination of HOG with SVM and the Haar cascade does not improve the accuracy of the model. As is shown in Table 5, the combination of HOG with SVM for specifying the size of faces and the proposed CNN model outperforms other combinations and the state-of-the-art face detection algorithm presented in OpenCV. In Figs. 11 and 12, we demonstrate absolute differences (ADs) and the mean absolute error (MAE) for nine groups of 109 images each, which are sorted by ground truth counts values from the smallest to the largest. In contrast to the DNN face detector which does not fully comply with the density changes, our proposed method has a uniform error in various densities.

Table 4. Performance comparison on the Pointing’04 (indicated by P) and FEI (indicated by F) datasets.

Techniques	Recall(P)	Recall(F)	Average
Adaboost	0.4439	0.8514	0.647
Simple	0.4224	0.3214	0.3719
Head detector	0.5584	0.8971	0.7277
Object detector	0.3656	0.8036	0.5846
Face-p146	0.7887	0.9939	0.8913
FFD	0.9969	0.9957	0.9963
HPM	0.8379	0.9792	0.9085
DPM	0.9977	0.9257	0.9617
IQA	0.7189	0.9368	0.8278
Face++	0.3472	0.8157	0.5814
MMOD	0.5337	N/A	N/A
NPD	0.9547	0.9932	0.9739
SSH	0.9984	0.9921	0.9952
IOP	0.9416	0.9882	0.9649
DNN	0.9884	N/A	N/A
Proposed	0.9976	0.9828	0.9902

5. Conclusion

In this paper, we propose an approach for face detection-based crowd estimation under significant occlusion and head pose variations. Several effective strategies have been proposed to increase the robustness of the CNN model. By randomly segmenting crowd images to generate face datasets and using them in the training phase of the CNN algorithm, the neural network

Table 5. Comparison of different methods.

Methods	SAD	MAE
Haar cascade + Proposed model	1237	11.34
HOG with SVM + Proposed model	676	6.2
HOG with SVM + HAAR + Proposed model	873	8
HOG with SVM + DNN	1754	16.09

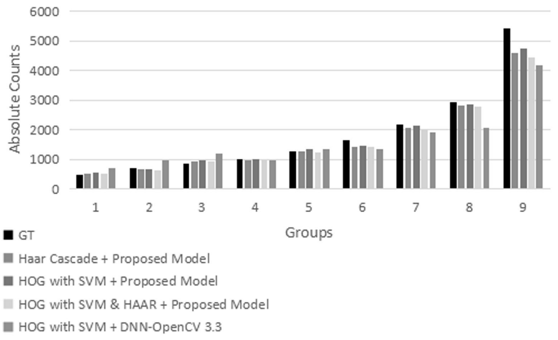


Fig. 11. Comparison of the absolute counts (histogram).

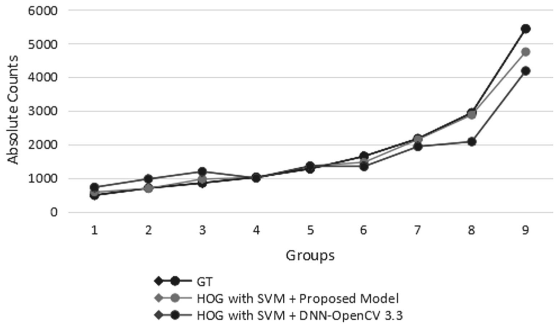


Fig. 12. Comparison of the absolute counts.

model can learn various features of faces and mitigate the negative effects of occlusion and head pose variations in the face detection approach.

The proposed method combines the well-known traditional machine learning algorithm with the state-of-the-art CNN algorithm to predict faces in crowded scenes. A new sliding window approach is implemented using the combination mentioned above. Combining the GHE-based approach with several different strategies allows the model to reconstruct faces from images with various lighting. Utilizing a network based on the VGGNet architecture with fewer convolutional layers, the proposed algorithm outperforms various face detection algorithms on the most challenging data sets. Our model is mostly trained on small face images (with sizes lower than 50×50 pixels) and achieves the best results on images consisting of faces with sizes in the range of 18–43 pixels. Since our model is robust predominantly on the detecting of tiny faces, it can be improved by training different models on various face scales. Therefore, in future work, addressing the scalability of the proposed method we will improve the detection performances on images consisting of larger face images.

Acknowledgment

This work was supported by the *ECHO* project, funded through the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 830943.

References

Buades, A., Coll, B. and Morel, J.-M. (2011). Non-local means denoising, *Image Processing On Line* **1**: 208–212, DOI: 10.5201/ipol.2011.bcm_nlm.

Cevikalp, H. and Triggs, B. (2012). Efficient object detection using cascades of nearest convex model classifiers, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, USA*, pp. 3138–3145.

Chan, A.B., Liang, Z.-S.J. and Vasconcelos, N. (2008). Privacy preserving crowd monitoring: Counting people without people models or tracking, *2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA*, pp. 1–7.

Chen, D., Ren, S., Wei, Y., Cao, X. and Sun, J. (2014). Joint cascade face detection and alignment, *in* D. Fleet *et al.*, (Eds), *Computer Vision—ECCV 2014*, Springer, Cham, pp. 109–122.

Chen, J., Deng, Y., Bai, G. and Su, G. (2015). Face image quality assessment based on learning to rank, *IEEE Signal Processing Letters* **22**(1): 90–94.

Chen, K., Loy, C.C., Gong, S. and Xiang, T. (2012). Feature mining for localised crowd counting, *British Machine Vision Conference (BMVC), Surrey, UK*.

Conte, D., Foggia, P., Percannella, G., Tufano, F. and Vento, M. (2010). A method for counting people in crowded scenes, *7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, USA*, pp. 225–232.

Davies, A.C., Yin, J.H. and Velastin, S.A. (1995). Crowd monitoring using image processing, *Electronics Communication Engineering Journal* **7**(1): 37–47.

Face++ (2015). Face detection software, `http://www.faceplusplus.com`.

Ferryman, J. and Ellis, A.-L. (2010). PETS2010: Dataset and challenge, *7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, USA*, Vol. 1, pp. 143–150.

Fradi, H. and Dugelay, J. (2012). Low level crowd analysis using frame-wise normalized feature for people counting, *2012 IEEE International Workshop on Information Forensics and Security (WIFS), Costa Adeje, Spain*, pp. 246–251.

Ghiasi, G. and Fowlkes, C. (2014). Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, USA*, pp. 1899–1906.

Gourier, N., Hall, D. and Crowley, J.L. (2004). Estimating face orientation from robust detection of salient facial features, *ICPR International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK*, pp. 17–25.

Han, D., Kim, J., Ju, J., Lee, I., Cha, J. and Kim, J. (2014). Efficient and fast multi-view face detection based on feature transformation, *16th International Conference on Advanced Communication Technology, Pyeongchang, Korea (South)*, pp. 682–686.

Idrees, H., Saleemi, I., Seibert, C. and Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, USA*, pp. 2547–2554.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift, *CoRR:* abs/1502.03167, http://arxiv.org/abs/1502.03167.

Jiang, H. and Learned-Miller, E. (2017). Face detection with the faster r-CNN, *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, USA*, pp. 650–657.

Jones, M. and Viola, P. (2003). Fast multi-view face detection, *Technical Report TR2003-96*, MERL—Mitsubishi Electric Research Laboratories, Cambridge, http://www.merl.com/publications/TR2003-96/.

Keras (2015). Keras extras code, https://github.com/keras-team/keras.

Khatoon, R., Saqlain, S.M. and Bibi, S. (2012). A robust and enhanced approach for human detection in crowd, *2012 15th International Multitopic Conference (INMIC), Islamabad, Pakistan*, pp. 215–221.

Kian Ara, R. and Matiolanski, A. (2019). *AGH Crowd Density Estimation Database (ACD)*, http://kt.agh.edu.pl/matiolanski/CrowdDensityEstimationDatabase/.

King, D.E. (2009). Dlib-ml: A machine learning toolkit, *Journal of Machine Learning Research* **10**: 1755–1758.

King, D.E. (2015). Max-margin object detection, *CoRR*: abs/1502.00046, http://arxiv.org/abs/1502.00046.

Kong, D., Gray, D. and Tao, H. (2005). Counting pedestrians in crowds using viewpoint invariant training, *BMVC 2005—Proceedings of the British Machine Vision Conference, Oxford, UK*.

Kotan, M., Öz, C. and Kahraman, A. (2021). A linearization-based hybrid approach for 3D reconstruction of objects in a single image, *International Journal of Applied Mathematics and Computer Science* **31**(3): 501–513, DOI: 10.34768/amcs-2021-0034.

Li, J. and Zhang, Y. (2013). Learning surf cascade for fast and accurate object detection, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Portland, USA*, pp. 3468–3475.

Li, S.Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H. and Shum, H. (2002). Statistical learning of multi-view face detection, *in* A. Heyden *et al.* (Eds), *Computer Vision—ECCV 2002*, Springer, Berlin, pp. 67–81.

Liao, S., Jain, A.K. and Li, S.Z. (2016). A fast and accurate unconstrained face detector, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(2): 211–223.

Liu, S., Dong, Y., Liu, W. and Zhao, J. (2012). Multi-view face detection based on cascade classifier and skin color, *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, Hangzhou, China*, Vol. 01, pp. 56–60.

Ma, R., Li, L., Huang, W. and Tian, Q. (2004). On pixel count based crowd density estimation for visual surveillance, *IEEE Conference on Cybernetics and Intelligent Systems, 2004, Singapore*, Vol. 1, pp. 170–173 vol.1.

Mahbub, U., Patel, V., Chandre, D., Barbello, B. and Chellappa, R. (2016). Partial face detection for continuous authentication, *CoRR*: abs/1603.09364, https://arxiv.org/abs/1603.09364.

Marin-Jimenez, M., Zisserman, A. and Ferrari, V. (2011). "Here's looking at you, kid". Detecting people looking at each other in videos, *Proceedings of the British Machine Vision Conference, Dundee, UK*, pp. 1–12, DOI: 10.5244/C.25.22.

Najibi, M., Samangouei, P., Chellappa, R. and Davis, L.S. (2017). SSH: Single stage headless face detector, *2017 IEEE International Conference on Computer Vision (ICCV), Los Alamitos, USA*, pp. 4885–4894.

Open CV (2019). *Open Source Computer Vision Library*, https://github.com/opencv/opencv.

Opitz, M., Waltner, G., Poier, G., Possegger, H. and Bischof, H. (2016). Grid loss: Detecting occluded faces, *CoRR*: abs/1609.00129, http://arxiv.org/abs/1609.00129.

Orozco, J., Martineza, B. and Pantic, M. (2015). Empirical analysis of cascade deformable models for multi-view face detection, *Image and Vision Computing* **42**: 47–61.

Pai, Y.-T., Ruan, S.-J., Shie, M.-C. and Liu, Y.-C. (2006). A simple and accurate color face detection algorithm in complex background, *2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada*, Vol. 2006, pp. 1545–1548.

Ren, S., He, K., Girshick, R.B. and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks, *CoRR*: abs/1506.01497, http://arxiv.org/abs/1506.01497.

Rosebrock, A. (2021). MiniVGGNet: Going deeper with CNNs, https://pyimagesearch.com/2021/05/22/minivggnet-going-deeper-with-cnns/.

Saleh, S.A.M., Suandi, S.A. and Ibrahim, H. (2015). Recent survey on crowd density estimation and counting for visual surveillance, *Engineering Applications of Artificial Intelligence* **41**: 103–114.

Shapiro, L.G. and Stockman, G.C. (2001). *Computer Vision*, Prentice Hall PTR, Upper Saddle River, pp. 137–150.

Sheng-Fuu, L., Jaw-Yeh, C. and Hung-Xin, C. (2001). Estimation of number of people in crowded scenes using perspective transformation, *IEEE Transactions on Systems, Man, and Cybernetics A: Systems and Humans* **31**(6): 645–654.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv*: 1409.1556, https://arxiv.org/abs/1409.1556.

Sun, X., Wu, P. and Hoi, S. (2017). Face detection using deep learning: An improved faster RCNN approach, *Neurocomputing* **299**: 42–50.

Thomaz, C. and Giraldi, G. (2010). A new ranking method for principal components analysis and its application to face image analysis, *Image and Vision Computing* **28**(6): 902–913.

Tsai, Y.-H., Lee, Y.-C., Ding, J.-J., Y. Chang, R. and Hsu, M.-C. (2018). Robust in-plane and out-of-plane face detection algorithm using frontal face detector and symmetry extension, *Image and Vision Computing* **78**: 26–41.

Viola, P. and Jones, M.J. (2004). Robust real-time face detection, *International Journal of Computer Vision* **57**(2): 137–154, DOI: 10.1023/B:VISI.0000013087.49260.fb.

Wan, S., Chen, Z., Zhang, T., Zhang, B. and Wong, K. (2016). Bootstrapping face detection with hard negative examples, *CoRR*: abs/1608.02236, http://arxiv.org/abs/1608.02236.

Yang, B., Yan, J., Lei, Z. and Li, S.Z. (2014). Aggregate channel features for multi-view face detection, *CoRR*: abs/1407.4023, http://arxiv.org/abs/1407.4023.

Yang, S., Luo, P., Loy, C.C. and Tang, X. (2015). From facial parts responses to face detection: A deep learning approach, *2015 IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile,* pp. 3676–3684.

You-jia, F. and Jian-wei, L. (2010). Rotation invariant multi-view color face detection based on skin color and ADaBOost algorithm, *2015 IEEE International Conference on Computer Vision (ICCV), Wuhan, China.*

Zhao, T., Nevatia, R. and Wu, B. (2008). Segmentation and tracking of multiple humans in crowded environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(7): 1198–1211.

Zhu, Q., Yeh, M.-C., Cheng, K.-T.T. and Avidan, S. (2006). Fast human detection using a cascade of histograms of oriented gradients, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, New York, USA,* pp. 1491–1498.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, USA,* pp. 2879–2886.

Zitouni, M.S. and Śluzek, A. (2022). A data association model for analysis of crowd structure, *International Journal of Applied Mathematics and Computer Science* **32**(1): 81–94, DOI: 10.34768/amcs-2022-0007.

Rouhollah Kian Ara received his BS degree in electrical and telecommunications engineering from Urmia Azad University, Iran, in 2010 and his MS degree in electrical and electronics engineering from Çukurova University, Adana, Turkey, in 2013. He is a PhD candidate in electronics and telecommunications engineering at the AGH University of Science and Technology, Cracow, Poland. He has been a research assistant at the AGH University since 2015. His research interests are artificial intelligence, machine learning and deep learning.

Andrzej Matiolanski received his BS and MS degrees in applied computer science from the AGH University of Science and Technology, Cracow, Poland, in 2010 and his PhD degree in telecommunications from the same university in 2017. From 2013 to 2017, he was a research assistant with the Department of Telecommunications. Since 2017, he has been an assistant professor with the Institute of Telecommunications, AGH University of Science and Technology in Cracow, Poland. He is the author of more than 20 publications (books and scientific papers). Andrzej Matiolanski is a vice-leader of the image data processing group at the Institute of Telecommunications. His research interests include computer vision, machine learning and algorithms.

Michał Grega received his MEng and PhD degrees in telecommunications from the AGH, University of Science and Technology, Kraków, in 2006 and 2011, respectively. In 2012 he completed a postgraduate course in project management at the Kraków University of Economics, receiving an IPMA D certificate. In 2012 he was appointed a visiting scholar at Berkeley University, CA, USA. Michał Grega is a co-author of over 50 research papers.

Andrzej Dziech holds a PhD degree in telecommunications from the Electrotechnical Institute in St Petersburg, Russia. He also holds a DSc degree in telecommunications from the Poznań University of Technology, Poland. His fields of interest are related to digital communication, image and data processing, intelligent monitoring, security systems, information and coding theory, random signals, computer communications networks and signal processing. He has worked at a number of foreign universities, in particular as a visiting professor at the University of Wuppertal in Germany (2001–2005). He is the author and a co-author of six books and 215 publications. He has supervised 19 PhD students. He has been honored four times by the Ministry of Science and Education of Poland for his research achievements, and recently by Prime Minister. He has been the coordinator of many national and 6 international projects, including the FP7 integrated project *INDECT* and the Operational Programme project *INSIGMA*.

Remigiusz Baran holds a PhD degree in telecommunications from the Faculty of Electrical, Control, Electronic and Computer Engineering, AGH University of Science and Technology in Kraków. He works as an assistant professor at the Kielce University of Technology. He is the author or a co-author of over 60 publications in the field of digital signal processing, in general focusing on image compression and visual content analysis of images and video sequences. His other research areas are feature-based object detection and recognition, as well as microprocessor technology and embedded systems.

Received: 12 May 2022
Revised: 8 August 2022
Re-revised: 27 October 2022
Accepted: 9 November 2022